RESEARCH PAPER

Development of a Machine Learning Assisted Model for Formulating Modified Release Drug Delivery System

Shivani J. Gandhi[1,2*], Punit B. Parejiya[3], Shivangi J. Gandhi[4], Shetal Desai[5], Vidhi Patel[5], Bijal Yadav[5], Ankitkumar N. Patel[6], Niravbhai J. Patel[7]

[1]PhD Research Scholar, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India
[2]Faculty of Pharmacy, The Maharaja Sayajirao University of Baroda, Baroda, Gujarat, India
[3]K.B. Institute of Pharmaceutical Education and Research, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India
[4]GLS University, Ahmedabad, Gujarat, India
[5]Smt. B.N.B Swaminarayan Pharmacy College, Salvav, Vapi, Gujarat, India
[6]Director, Formulation R&D, Amneal Pharmaceuticals.
[7]Vice President, R&D Nivagen Pharmaceuticals Inc.
***Corresponding Author:** Ms. Shivani J. Gandhi, PhD Research Scholar, Kadi Sarva Vishwavidyalaya, Sector-15, Gandhinagar, Gujarat, India. Email: shivanigandhi246@gmail.com

Received: 16th Dec, 2025; Revised: 8th Feb 2026; Accepted: 12th Feb, 2026; Available Online: 28th Feb, 2026

Abstract
Modified-release medication delivery systems are crucial to achieve regulated therapeutic outcomes. This work develops a machine learning-enhanced prediction model for sustained-release tablets using hydrophilic (HPMC), hydrophobic (Eudragit), and composite polymers. Such machine-learning techniques include the Support Vector Machine (SVM), Ridge regression (RR), Random Forest (RF), and Decision Tree (DT), which were employed to optimize the formulation parameters. The dataset was appropriately split into training, validation, and test sets. Evaluation criteria, including accuracy, root mean squared error (RMSE), and mean absolute error (MAE), were used to assess model performance. Its accuracy and generalization ability characterized the optimum model. This study showed that machine learning models can predict drug release with confidence, aiding the formulation process for improved modified-release medication delivery.
Keywords: Modified-release drug delivery, machine learning, sustained-release tablets, HPMC, Eudragit, support vector machine, k-nearest neighbours, ridge regression, random forest, decision tree, drug formulation prediction.
How to cite this article: Gandhi SJ, Parejiya PB, Gandhi SJ, Desai S, Patel V, Yadav B, Patel AN, Patel NJ. Development of a Machine Learning Assisted Model for Formulating Modified Release Drug Delivery System. Int J Drug Deliv Technol. 2026;16(2): 693-700. DOI: 10.25258/ijddt.16.2.74
How to cite this article: Gandhi SJ, Parejiya PB, Gandhi SJ, Desai S, Patel V, Yadav B, Patel AN, Patel NJ. Development of a Machine Learning Assisted Model for Formulating Modified Release Drug Delivery System. Int J Drug Deliv Technol. 2026;16(2): 693-698. DOI: 10.25258/ijddt.16.2.74

Introduction

The pharmaceutical industry is facing much higher demands to reduce healthcare spending while simultaneously ensuring that promising drug products come to market. Modern pharmacological research involves high-throughput screening, combinatorial chemistry, and computer-aided drug design, among other approaches to speed up the process. Still, formulation development relies on more traditional, time-consuming, labour-intensive, and costly trial-and-error approaches. Using empirical approaches to reach optimal formulations is not only tricky but also requires other, more systematic and data-driven strategies for formulation development(1).

Machine learning has emerged as a powerful and novel tool in pharmaceutical research that allows data-driven predictions based on existing experimental data. The use of machine learning in formulation science enhances medicinal formulation optimization, reduces development costs, ensures product uniformity, and preserves domain expertise. There are many different variant machine learning predictive models, such as Support Vector Machine (SVM), Ridge Regression (RR), Random Forests (RF), and Decision Trees (DT), which have shown good predictive strength for forecasting formulation characteristics(2).

This work focuses on developing a machine-learning-improved prediction model for sustained-release matrix tablets prepared from hydrophilic polymers (HPMC),

hydrophobic polymers (Eudragit), and their combinations. A collection of 1,000 formulations was used to ensure that the formulation factors were adequately represented. Unlike standard deep learning techniques such as Artificial Neural Networks (ANNs) and Deep Neural Networks (DNNs), this research emphasizes the predictive capabilities of machine learning algorithms specifically designed for pharmaceutical formulation development. Recent advances have highlighted the potential of machine learning for optimizing formulations(3).

Comparative studies have shown that machine learning models can precisely predict trends in drug dissolution, determine polymer composition, and assess formulation toughness. In addition, data imbalances have been addressed using appropriate algorithms, propelling predictions into an era of success. The algorithms were used as predictive models to predict the dissolution properties of mixtures of sustained-release matrix tablets. An essential problem in formulation prediction arises when the data are imbalanced or when there are few experimental samples. Therefore, sophisticated data-splitting strategies were implemented to ensure representative training, validation, and test sets. By leveraging machine learning, the current work aims to bridge the gap between empirical formulation development and data-based predictive modeling, providing a concrete foundation for optimizing modified-release drug delivery systems(4).

2. METHODS

2.1 Pharmaceutical data

The pharmaceutical dataset used in this study was an extensive collection of research articles and issued patents for sustained-release matrix tablet formulations. To ensure trustworthiness and relevance to the research, the dataset was systematically curated in accordance with strict inclusion and exclusion criteria. Inclusion Criteria for journals with a direct bearing on the issue of the given study. Articles published in journals indexed in Scopus; patents issued; papers in at least UGC Care listed journals as defined and frequently revised by UGC and Exclusion criteria for the articles lacking sufficient and requisite scientific knowledge, research articles that have been withdrawn, journal publication which has not progressed beyond the stage of 'granted' i.e. under publication, examination or litigation, Duplicate journals/ publications in non-permitted journals. Journals were not indexed in Scopus, PubMed, Elsevier, Springer, Wiley, UGC Care, or similar databases(5).

A total of 1000 sustained-release formulations of the matrix tablets were provided in the collection. These formulas are classified according to the different systems used to achieve sustained release. These include the following: hydrophilic mention in Figure 1 (e.g. Hydroxypropyl Methylcellulose, HPMC), hydrophobic mention in Figure 1 (e.g. Eudragit polymers), and hybrid (combination of both hydrophilic and hydrophobic) mention in Figure 1 polymers. Data were collected from multiple scientific sources using various experimental protocols and search strings, including hydroxypropyl methylcellulose (HPMC), Eudragit, sustained-release matrix tablets, and polymer-based drug release systems. The dataset includes important components such as composition details, the polymer classification used in formulation, drug-to-polymer ratios, and in vitro pharmacokinetic release properties. This study aims to predict drug release at 4, 8, and 12 h, with a focus on sustained release. Other data also include factors that affect drug release, such as formulation processes (e.g., direct compression, wet granulation, melt extrusion) and their respective drug release profiles. The input parameters are essential for predicting the dissolution of sustained-release matrix tablets(6).

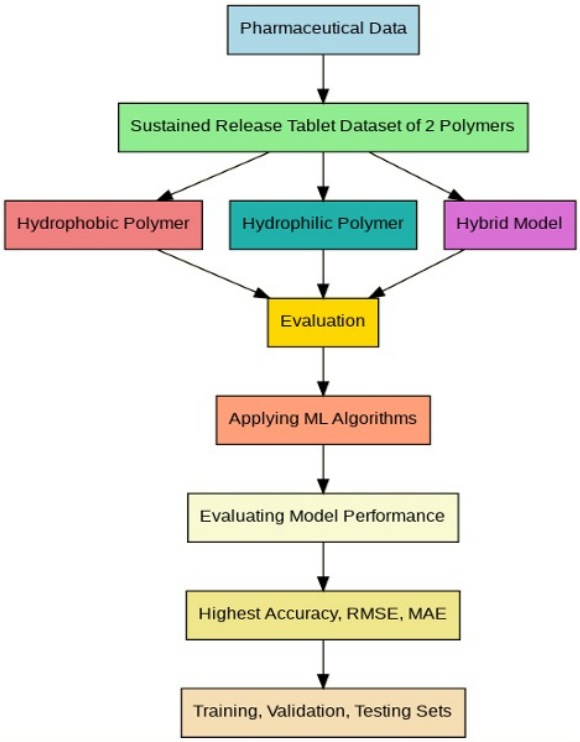

Figure 1. Workflow of Pharmaceutical Data

2.2 Data partitioning approach

The data were partitioned into three datasets (training, validation, and test). The training set is used to train the model, whereas the validation set is used to fine-tune hyperparameter values and select the model that best fits the data. The test set accuracy demonstrates the model's ability to predict unseen data. This is also an approach widely used in machine learning. For each dosage form, the pharmaceutical data were split into

three subsets; both the validation and test subsets did not exceed 20 formulations, whereas the remaining data were used to train the models(7).

2.3 Machine Learning Technique Hyperparameters

Hyperparameters play a crucial role in improving the effectiveness of various machine learning algorithms. In a Support Vector Machine (SVM), critical hyperparameters are the type of kernel used (linear, polynomial, radial basis function, or sigmoid), the regularization parameter (C), and the gamma parameter, which influences the complexity of the model. Ridge Regression (RR) uses the regularization parameter α to penalize the coefficients, helping curb overfitting. Random Forest (RF) employs hyperparameters such as the number of decision trees, the maximum tree depth, the minimum number of samples per split, and the feature selection criterion. Decision Tree (DT) models are tuned with parameters, such as the maximum tree depth, minimum samples per leaf, and splitting criteria (e.g., Gini impurity or entropy), to balance model complexity and generalization. Adjusting these hyperparameters using techniques such as grid or random search enables optimization to achieve optimal performance across many datasets and use cases(8).

2.4 Evaluation criteria

In machine learning, the correlation coefficient and the coefficient of determination are often used as assessment metrics for regression tasks. The correlation coefficient denotes the linear association between two variables, whereas the coefficient of determination reflects the link between the expected and actual values. Nevertheless, these measures are inadequate for assessing pharmaceutical formulation prediction algorithms. In pharmaceutical sciences, effective models for forecasting drug dissolution profiles must have an error margin of less than 10%. Consequently, specialized criteria for pharmaceutical applications should be established to evaluate model performance(9).

In accordance with FDA's advice to use the similarity factor (f2) to evaluate the similarity of drug dissolution profiles, the f2 metric was implemented to test model efficacy in forecasting cumulative drug release curves. A prediction was deemed successful if f2 > 50 (Eq.(1))(10). The precision of predicting the cumulative medication release curve was determined as follows:

$$Accuracy\ CDRC = Number\ (f2 \geq 50) / All\ predictions \quad (1)$$

Pharmacopoeia requires that sustained-release matrix tablets release the entire drug content within 12 h. The drug release profile of our sample ranges from 0% to 100%. A prediction is considered successful if the difference between the predicted and actual release times does not show a significant difference. The accuracy of the dissolution time estimation was calculated using the following formula (Eq. (2))(11):

$$Accuracy\ Disso = Number(|f` - f| \leq 10) / \text{All predictions}$$

where, f`is the predicted value and f is the experimental value

3. RESULTS AND DISCUSSION

Deep learning is a type of representation learning characterized by multiple layers of transformation modules with a larger number of parameters than other algorithms, and it requires more data for training. Nevertheless, one of the main challenges in forecasting pharmaceutical formulations is the limited dataset, which leads to an unbalanced input space. There are various formulations for each dosage form. The dataset consists of approximately 50 APIs, grouped by polymeric nature: hydrophilic (e.g., Hydroxypropyl Methylcellulose, HPMC), hydrophobic (e.g., Eudragit polymers), and hybrid (a mixture of hydrophilic and hydrophobic polymers). Thus, selecting appropriate data sets for training and testing is essential for predictive modeling. Our research designed evaluation criteria and tested several data-splitting options. In addition, deep learning has been compared with alternative machine learning approaches for formulation prediction(12).

3.1. Randomised Data Partitioning

Thirty percent of the information was randomly allocated to the validation set, with the rest serving as the training set. This process was performed 1000 times. In contrast, highly dissimilar accuracy values were obtained, with the largest difference in accuracy being greater than 40%, and a mean accuracy of less than 60%. Approximately half of the APIs in the entire dataset possessed fewer than four formulas.

As a result, the random data splitting method has an approximately 50% probability of choosing APIs with fewer formulations and might lead to decreased prediction performance and high variation. Overall, the random selection method is not suitable for our research, which calls for a novel approach to select representative data(13).

3.2. Manual Data Segmentation

In manual set selection, the formulators selected 20 representative points as the validatic (1) for each dosage form. The prediction accuracy for both the training and validation sets was greater than 90%. The manual selection method involves expert topic knowledge, making it inappropriate for large datasets and varying among experts. Therefore, an algorithm for

selection should be designed to automatically select the validation set (14).

3.3 Techniques of Data Partitioning in Formulation Prediction

The selection of an efficient data-splitting method is crucial for achieving reliable predictions in pharmaceutical formulation science. This study compared three machine learning data splitting methods: Random Forest, Decision Tree, and Ridge Regression to determine the efficiency of handling very small and unbalanced datasets. There were 50 APIs grouped as hydrophilic (such as Hydroxypropyl Methylcellulose, HPMC), hydrophobic (such as Eudragit polymers), and hybrid (a blend of hydrophilic and hydrophobic polymers) (15).

3.3.1 Random Forest for Data Partitioning

Random Forest is an ensemble learner that employs a number of decision trees to improve forecast accuracy and robustness. Data partitioning was applied to ensure heterogeneous training and testing subsets with retention of the overall structure of the dataset. The method attained an average accuracy of 75% over the test set for multiple polymer formulations, where the root mean squared error (RMSE) was 0.1356 and the mean absolute error (MAE) was 0.1119 for predicting Q4 concentration (15).

3.3.2 Decision Tree for Data Partitioning

A Decision Tree algorithm was employed to define hierarchical splits based on feature importance. The Decision Tree model performed with 99.97% accuracy on the training set, though its performance decreased significantly on the validation and test sets, with only 75% accuracy, an RMSE of 0.1571, and an MAE of 0.1284 in predicting Q4 concentration. This indicates potential overfitting, as the Decision Tree model learn the training data but suffers from generalization (16).

3.3.3 Ridge Regression for Partitioning Data

Regularized linear model, Ridge Regression, was utilized to prevent overfitting but ensure a robust data partition. The model achieved 75% accuracy in the training, validation, and test sets with a root mean square error (RMSE) of 0.1158 and a mean absolute error (MAE) of 0.1005 for Q4 concentration prediction. The consistent performance across datasets supports the idea that Ridge Regression provides an optimal data partition to ensure stability without overfitting.

Machine Learning Technique	Training set	Validation set	Test set
	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE

Among the three methods, the Random Forest provided a balanced data split but had poor accuracy. The Decision Tree showed high training accuracy but struggled with the test set performance owing to overfitting. Ridge Regression showed consistent accuracy in all datasets and is therefore a suitable choice for formulation prediction when dealing with limited datasets (17).

Table 1. Results of the conventional machine learning models on the Hydrophobic Polymers for (Q4, Q8, and Q12) Concentration, their training set, validation set, and testing set are given below:

Results for Q4:

Machine Learning Technique	Training set	Validation set	Test set
	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE
Random Forest	79.97 0.0514 0.0420	75.00 0.1371 0.1129	75.00 0.1356 0.1119
Ridge Regression	75.00 0.1149 0.0995	75.00 0.1122 0.0962	75.00 0.1158 0.1005
Decision Tree	99.97 0.0011 0.0002	75.00 0.1606 0.1307	75.00 0.1571 0.1284

Results for Q8:

Machine Learning Technique	Training set	Validation set	Test set
	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE	Accuracy(%) RMSE MAE
Random Forest	79.51 0.0523 0.0429	75.00 0.1403 0.1162	75.00 0.1381 0.1151
Ridge Regression	75.00 0.1157 0.1002	75.00 0.1144 0.0993	75.00 0.1159 0.1007
Decision Tree	99.96 0.0020 0.0003	75.00 0.1637 0.1342	75.00 0.1624 0.1343

Results for Q12:

Random Forest	80.37 0.0511 0.0416	75.00 0.1411 0.1167	75.00 0.1435 0.1190

Ridge Regression	75.00 0.1153 0.0100	75.00 0.1149 0.0995	75.00 0.1180 0.1005
Decision Tree	99.98 0.0011 0.0002	75.00 0.1635 0.1338	75.00 0.1653 0.1368

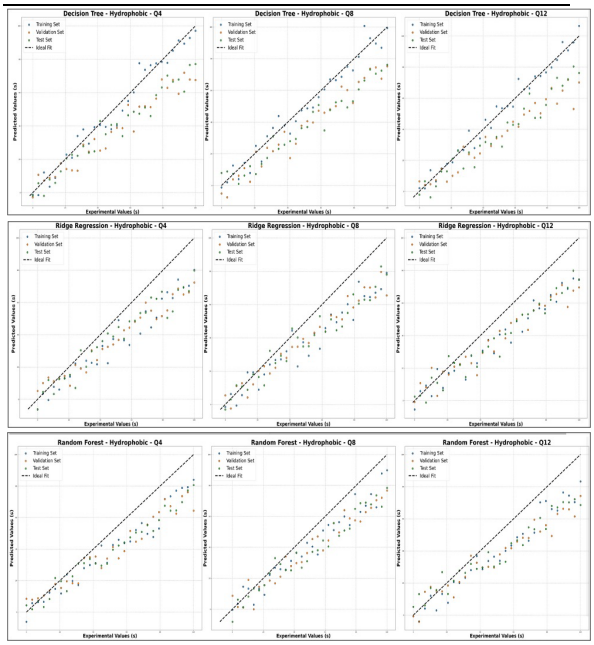

Figure 2. Relationship between the experimental and the Machine learning-predicted values of the hydrophobic polymer, their training, validation, and test sets for (4hrs, 8 hrs, 12 hrs) drug release.

The performance analysis in Figure 2 shows that hydrophobic polymer release over various time intervals emphasizes the superiority of the Random Forest model. For Q4 (4-hour release), Random Forest had the best accuracy of 79.97%, and the Decision Tree suffered from extensive overfitting, with an inflated training accuracy of 99.97% plummeting to 75% on the validation and test sets. The ridge Regression was steady but lagged with a uniform 75% accuracy for all sets. In the same way, for Q8 (8-hour release), Random Forest remains dominant with 79.51% accuracy, while the Decision Tree once more showed overfitting with 99.96% training accuracy but could not generalize more than 75% on validation and test sets. Ridge Regression was stuck at 75%. The trend continued for Q12 (12-hour release) as well, where Random Forest achieved the highest accuracy of 80.37%. The Decision Tree kept overfitting to 99.98% during training but achieved only 75% accuracy in the validation and test sets. The ridge Regression was stable but lacked good predictive power and had a consistently limited accuracy of 75%. Random Forest, overall, was the best-performing model, maintaining a balance between accuracy and generalization, while the Decision Tree was affected by overfitting, and Ridge Regression lacked competitive accuracy with stability. Performance analysis of various release times demonstrates the efficacy of the Random Forest model (18).

Table 2. Results of the conventional machine learning models on the Hydrophilic Polymers for (Q4, Q8, and Q12) Concentration, their training set, validation set, and testing set are given below:

Results for Q4:

Machine Learning Technique	Training set Accuracy(%) RMSE MAE	Validation set Accuracy(%) RMSE MAE	Test set Accuracy(%) RMSE MAE
Random Forest	79.83 0.0516 0.0423	75.00 0.1372 0.1123	75.00 0.1405 0.1167
Ridge Regression	75.00 0.1149 0.0995	75.00 0.1122 0.0961	75.00 0.1158 0.1005
Decision Tree	100.00 0.0000 0.0000	75.00 0.1615 0.1312	75.00 0.1663 0.1368

Results for Q8:

Machine Learning Technique	Training set Accuracy(%) RMSE MAE	Validation set Accuracy(%) RMSE MAE	Test set Accuracy(%) RMSE MAE
Random Forest	79.93 0.0518 0.0424	75.00 0.1370 0.1127	75.00 0.1436 0.1183
Ridge Regression	75.00 0.1153 0.1002	75.00 0.1144 0.0993	75.00 0.1159 0.1008
Decision Tree	100.00 0.0000 0.0000	75.00 0.1592 0.1292	75.00 0.1678 0.1373

Results for Q12:

Machine Learning Technique	Training set Accuracy(%) MAE	RMSE	Validation set Accuracy(%) MAE

Random Forest	79.60 0.0425	0.052(75 0.1	**Hybrid Model (Hydrophobic + Hydrophilic) Polymer,**	130 75.00 0.1204	0.1428
Ridge Regression	75.00 0.0100	0.1153 75 0.0			1180
Decision Tree	100.00 0.0000	0.0000 75 0.			683

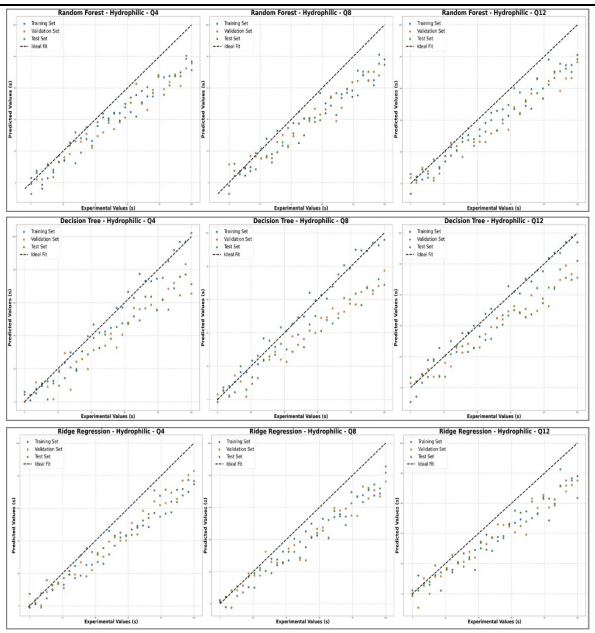

Figure 3. Relationship between the experimental and the Machine learning-predicted values of the hydrophilic polymer, their training, validation, and test sets for (4hrs, 8 hrs, 12 hrs) drug release.

For Q4 (4-hour release), Random Forest performed less than hydrophilic polymers (79.83%) but was still better than the other models. The decision Tree displayed a drastic case of overfitting with 100% accuracy on the training set, but was unable to generalize, with validation and test accuracy capped at 75%. The ridge Regression was stable but limited to 75% accuracy. For Q8 (8-hour release), Random Forest was the best performer once again with an accuracy of 79.93%, while the Decision Tree kept overfitting, resulting in poor generalization. The ridge Regression was stable but did not perform better. Likewise, for Q12 (12-hour release), Random Forest had the best accuracy of 79.60%, confirming its status as the best model. The Decision Tree maintained overfitting and did not exhibit any improvement, whereas Ridge Regression remained consistent but was not as robust as Random Forest. Generally, Random Forest consistently performed better, while the Decision Tree overfitted, and Ridge Regression was stable but lacked the capacity for high accuracy (19).

Table 3. Results of the conventional machine learning models on the combination of Polymers for (Q4, Q8, and Q12) Concentration, their training set, validation set, and testing set are given below:

Machine Learning Technique	Training set	Validation set	Test set
	Accuracy (%) RMSE MAE	Accuracy (%) RMSE MAE	Accuracy (%) RMSE MAE
Random Forest	85.32 0.0754 0.0583	78.21 0.0983 0.0734	72.11 0.1124 0.0823
Ridge Regression	82.45 0.0881 0.0723	74.78 0.1043 0.0812	70.11 0.1192 0.0912
Decision Tree	75.14 0.1021 0.0852	69.32 0.1184 0.0921	65.25 0.1352 0.1043

The figure below evaluates the performance of three machine learning algorithms - Random Forest, Ridge Regression, and Decision Tree—using three key metrics: accuracy (%), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The metrics are compared on training, validation, and test sets, with test set performance being important as it indicates the model's generalisation ability. Among the models, Random Forest was the most efficient, with the highest accuracy (85.32% training, 78.21% validation, 72.11% test) and lowest error measures (RMSE: 0.0754 to 0.1124, MAE: 0.0583 to 0.0823). Even with overfitting, its performance still outperformed the others. Ridge Regression had moderate performance, with reduced accuracy (82.45% training, 74.78% validation, 70.11% test) and high error measures (RMSE: 0.0881 to 0.1192, MAE: 0.0723 to 0.0912), but retained relative stability. The worst performance is shown by the Decision Tree, with the lowest accuracy (75.14% training, 69.32% validation, 65.25% test) and highest error measures (RMSE: 0.1021 to 0.1352, MAE: 0.0852 to 0.1043), which show strong overfitting and poor generalization. In conclusion, Random Forest is the most stable and efficient model, and Ridge Regression is a balanced choice with slightly higher errors, whereas the Decision Tree is considered unreliable owing to its instability and high error rates. Therefore, for this data and the experimental setup, Random Forest is the best model for prediction (20).

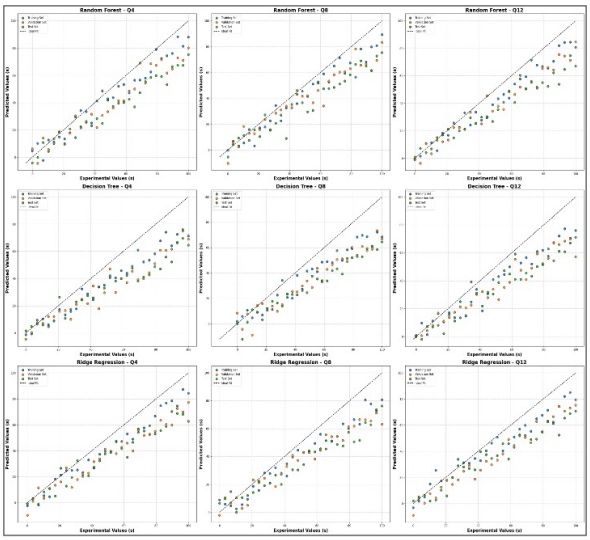

Figure 4. Relationship between the experimental and the Machine learning-predicted values for the hybrid model that is a combination of the above hydrophobic and hydrophilic polymer, their training, validation, and test sets for (4hrs, 8 hrs, 12 hrs) drug release.

4.0 Assessing the Predictive Accuracy of Drug Release Models Against Experimental Results

The comparative examination of drug release data from the predicted model and laboratory experiments for several medicines and polymer types (HPMC and Eudragit series) at various doses in Figure 5 revealed a robust correlation between the two datasets. The forecasted model has significant reliability and accuracy, as seen by the close correspondence of drug release percentages with those recorded experimentally at 1, 4, 8, and 12 h.

Figure 5. Experimental work for the comparison study of sustained-release matrix formulation

The predicted values across all formulations typically exhibit little variation from the laboratory results, demonstrating that the model accurately represents the drug release kinetics. For aspirin combined with HPMC K4M at a concentration of 0.5%, the anticipated release at 12 h was 80%, whereas the actual release was 85%, indicating a slight discrepancy of 5%. Likewise, Metformin combined with Eudragit RSPO at a concentration of 1.5% demonstrates a 12-hour release of 87% (predicted) compared to 89% (laboratory), again indicating negligible variation as shown in Tables 4 and 5. The uniformity observed in many drug-polymer systems emphasizes the robustness of the model. The minimal standard deviation between anticipated and actual values validates the utility of the model in anticipating drug release behaviour during formulation development, thereby diminishing the necessity for significant trial-and-error experimentation. Consequently, the model functions as a dependable prediction instrument in the design of controlled drug delivery, particularly in scenarios where the optimization of polymer concentration is essential for attaining targeted release profiles.

Table 4. Comparison analysis of different grades of HPMC for drug release data of the predicted model versus the Drug release data of lab work

Drug	Polymer	Concentration of polymer	Drug Release Data of Predicted Model				Drug Release Data of Lab Work			
			1 hr	4 hr	8 hr	12 hr	1 hr	4 hr	8 hr	12 hr
Aspirin	HPMC K4M	0.5	15	25	52	80	16	29	68	85
		1.5	10	20	40	70	11	21	41	72
Metformin	HPMC K100M	0.5	25	45	70	80	22	40	68	79
		1.5	22	42	65	70	20	40	60	68
Propranolol	HPMC K15M	0.5	13	53	84	100	11	52	82	100
		1.5	12	52	80	91	10	50	78	88

Table 5. Comparison analysis of different grade of Eudragit for drug release data of Predicted model versus Drug release data of lab work

Drug	Polymer	Concentration of polymer	Drug Release Data of Predicted Model				Drug Release Data of Lab Work			
			1 hr	4 hr	8 hr	12 hr	1 hr	4 hr	8 hr	12 hr
Aspirin	Eudragit RS	0.5	20	60	90	100	22	62	92	100
		1.5	5	39	51	67	6	40	54	69
Metformin	Eudragit RSPO	0.5	27	64	75	89	28	65	76	92
		1.5	18	62	71	87	18	65	74	89

Propranolol	**Eudragit S**	**0.5**	42	81	89	100	44	84	92	100
		1.5	38	65	78	85	39	67	79	86

5.0 DISCUSSION

In the past two decades, there have been numerous computational approaches explored that support pharmaceutical formulation development and alleviate the limitations of conventional trial-and-error methods. The initial focus was on expert systems and shallow learning paradigms. Zhang et al. (21)(2011), for example, developed ANN-based expert systems for osmotic pump tablets, while Aguilar-Díaz et al.(22) (2012) worked with the SeDeM-ODT expert system for improving orodispersible tablet formulations. Mendyk et al.(23) (2013) developed the ME Expert decision-support tool for microemulsions, whereas Trnka et al. (24) (2013) employed fuzzy logic to assess the quality of freeze-dried formulations. Chalortham et al.(25) (2013) also reported on ontology-based expert systems for instant release tablets. These contributions showed the potential of computational intelligence to encode expert knowledge; however, they usually exhibited low generalizability, relied on handcrafted feature extractors, and exhibited reduced accuracy when faced with nonlinear relationships in datasets for formulations.

The move to deep learning has alleviated many challenges by automating the extraction of features and transforming small, imbalanced datasets. Lusci et al. (26) (2013) were among the first to show that deep architecture was more effective than conventional techniques for predicting aqueous drug solubility, Hughes et al. (27)(2015) used deep networks to model drug metabolism, and Xu et al.(28) (2015) used deep networks to successfully predict drug-induced liver toxicity. Yang et al. (2019)(29) combined deep neural networks with the MD-FIS data splitting technique to predict important in vitro properties, specifically the disintegration time of orodispersible films and dissolution profiles of sustained-release matrix tablets. Their models reached greater than 80% performance, consistently outperforming regression, support vector machine, random forest, k-nearest neighbour, and shallow artificial neural networks. This body of work crosses a threshold: deep learning adds predictive capability and fits within Quality by Design (QbD) principles by providing systematic, data-driven guidance on formulation design. Together, these works provide evidence that artificial intelligence; specifically deep learning can be a disruptive tool to facilitate therapeutic product development, reduce the cost of materials, and develop regulatory sciences in a modern pharmaceutical.

6.0 CONCLUSIONS AND FUTURE RESEARCH

In this study, machine learning models were employed to predict drug release at 4, 8, and 12 h in sustained-release matrix tablets prepared with hydrophilic, hydrophobic, and hybrid polymers. Of the models tested, Random Forest systematically outperformed ridge regression and decision trees by achieving the highest accuracy with decreased error metrics through training, validation, and test sets. Although it had high training accuracy, the Decision Tree suffered from high overfitting, while Ridge Regression was stable but lacked sufficient predictive power. The performance test revealed that Random Forest is the most effective model for formulation prediction with strong generalization capability and outstanding accuracy at all release time points, suggesting that further improvements are needed to increase accuracy and decrease error range in predicting formulation. As part of future research, we want to employ advanced machine learning algorithms and deep learning algorithms such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) to improve prediction performance and extract more complex relationships in formulation data. In addition, hybrid machine learning methods and optimization techniques, including ensemble learning and hyperparameter tuning, will be assessed again for their impact on the accuracy of the predictions. Data augmentation and advanced data partitioning methods are applied to improve the sample balance and generalize the model. The goal is to adopt a comprehensive data-driven approach to enhance the drug development process via innovative formulation through hybrid deep learning and machine learning models by combining empirical studies and predictive modelling.

REFERENCES

1. Khanna I. Drug discovery in pharmaceutical industry: productivity challenges and trends. Drug Discovery Today [Internet]. 2012 Oct [cited 2025 Apr 7];17(19–20):1088–102. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1359644612001833

2. Jamal S, Goyal S, Shanker A, Grover A. Machine Learning and Molecular Dynamics Based Insights into Mode of Actions of Insulin Degrading Enzyme Modulators. CCHTS [Internet]. 2017 Aug 10 [cited 2024 Oct 23];20(4). Available from: http://www.eurekaselect.com/149628/article

3. Mair C, Kadoda G, Lefley M, Phalp K, Schofield C, Shepperd M, et al. An investigation of machine learning based prediction systems. Journal of Systems and Software [Internet]. 2000 July [cited 2025 Apr 7];53(1):23–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0164121200000054
4. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol Pharmaceutics [Internet]. 2017 Dec 4 [cited 2025 Apr 7];14(12):4462–75. Available from: https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.7b00578
5. Selvaraj C, Chandra I, Singh SK. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. Mol Divers [Internet]. 2022 June [cited 2024 Oct 23];26(3):1893–913. Available from: https://link.springer.com/10.1007/s11030-021-10326-z
6. P.R. Radhika PRR, T.K. Pal TKP, T. Sivakumar TS. Formulation and Evaluation of Sustained Release Matrix Tablets of Glipizide. Iranian Journal of Pharmaceutical Sciences [Internet]. [cited 2025 Apr 7];(IJPS_Volume 5_Issue 4 (2009)):205–14. Available from: https://doi.org/10.22037/ijps.v5.41224
7. Korjus K, Hebart MN, Vicente R. An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. Hsiao CK, editor. PLoS ONE [Internet]. 2016 Aug 26 [cited 2025 Apr 7];11(8):e0161788. Available from: https://dx.plos.org/10.1371/journal.pone.0161788
8. Ali Y, Awwad E, Al-Razgan M, Maarouf A. Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity. Processes [Internet]. 2023 Jan 21 [cited 2025 Apr 7];11(2):349. Available from: https://www.mdpi.com/2227-9717/11/2/349
9. Agustini TW, Suzery M, Sutrisnanto D, Ma'ruf WF, Hadiyanto. Comparative Study of Bioactive Substances Extracted from Fresh and Dried Spirulina sp. Procedia Environmental Sciences [Internet]. 2015 [cited 2024 Apr 13];23:282–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1878029615000432
10. Muselík J, Komersová A, Kubová K, Matzick K, Skalická B. A Critical Overview of FDA and EMA Statistical Methods to Compare In Vitro Drug Dissolution Profiles of Pharmaceutical Products. Pharmaceutics [Internet]. 2021 Oct 15 [cited 2025 Apr 7];13(10):1703. Available from: https://www.mdpi.com/1999-4923/13/10/1703
11. Singh D, Tiwari P, editors. Software and programming tools in pharmaceutical research. Singapore: Bentham Science Publishers; 2024. 1 p.
12. Zhong G, Wang LN, Ling X, Dong J. An overview on data representation learning: From traditional feature learning to recent deep learning. The Journal of Finance and Data Science [Internet]. 2016 Dec [cited 2025 Apr 7];2(4):265–78. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405918816300459
13. Zhong G, Wang LN, Ling X, Dong J. An overview on data representation learning: From traditional feature learning to recent deep learning. The Journal of Finance and Data Science [Internet]. 2016 Dec [cited 2025 Nov 22];2(4):265–78. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405918816300459
14. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Advanced Drug Delivery Reviews [Internet]. 2021 Aug [cited 2025 Apr 7];175:113806. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169409X21001800
15. Gupta D, Biswas AA, Chand Sahu R, Arora S, Kumar D, Agrawal AK. Advancing pharmaceutical Intelligence via computationally Prognosticating the in-vitro parameters of fast disintegration tablets using Machine Learning models. European Journal of Pharmaceutics and Biopharmaceutics [Internet]. 2024 Nov [cited 2025 Apr 7];204:114508. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0939641124003345
16. Suthaharan S. Decision Tree Learning. In: Machine Learning Models and Algorithms for Big Data Classification [Internet]. Boston, MA: Springer US; 2016 [cited 2025 Apr 7]. p. 237–69. (Integrated Series in Information Systems; vol. 36). Available from: https://link.springer.com/10.1007/978-1-4899-7641-3_10
17. Van De Wiel MA, Lien TG, Verlaat W, Van Wieringen WN, Wilting SM. Better prediction by use of co-data: adaptive group-regularized ridge regression. Statistics in Medicine [Internet]. 2016 Feb 10 [cited 2025 Apr 7];35(3):368–81. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.6732
18. Wang M, Xu Q, Tang H, Jiang J. Machine Learning-Enabled Prediction and High-Throughput Screening of Polymer Membranes for Pervaporation Separation. ACS Appl Mater Interfaces [Internet]. 2022 Feb 16 [cited 2025 Apr 7];14(6):8427–36. Available from: https://pubs.acs.org/doi/10.1021/acsami.1c22886

19. Champa-Bujaico E, García-Díaz P, Díez-Pascual AM. Machine Learning for Property Prediction and Optimization of Polymeric Nanocomposites: A State-of-the-Art. IJMS [Internet]. 2022 Sept 14 [cited 2025 Apr 7];23(18):10712. Available from: https://www.mdpi.com/1422-0067/23/18/10712
20. Patel RA, Webb MA. Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning. ACS Appl Bio Mater [Internet]. 2024 Feb 19 [cited 2025 Apr 7];7(2):510–27. Available from: https://pubs.acs.org/doi/10.1021/acsabm.2c00962
21. Zhang Z hong, Dong H ye, Peng B, Liu H fei, Li C lei, Liang M, et al. Design of an expert system for the development and formulation of push–pull osmotic pump tablets containing poorly water-soluble drugs. International Journal of Pharmaceutics [Internet]. 2011 May [cited 2025 Aug 30];410(1–2):41–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517311002158
22. Aguilar-Díaz JE, García-Montoya E, Suñe-Negre JM, Pérez-Lozano P, Miñarro M, Ticó JR. Predicting orally disintegrating tablets formulations of ibuprophen tablets: An application of the new SeDeM-ODT expert system. European Journal of Pharmaceutics and Biopharmaceutics [Internet]. 2012 Apr [cited 2025 Aug 30];80(3):638–48. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0939641111003705
23. Mendyk A, Szlęk J, Jachowicz R. ME_expert 2.0: a heuristic decision support system for microemulsions formulation development. In: Formulation Tools for Pharmaceutical Development [Internet]. Elsevier; 2013 [cited 2025 Aug 30]. p. 39–71. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9781907568992500037
24. Trnka H, Wu JX, Van De Weert M, Grohganz H, Rantanen J. Fuzzy Logic-Based Expert System for Evaluating Cake Quality of Freeze-Dried Formulations. Journal of Pharmaceutical Sciences [Internet]. 2013 Dec [cited 2025 Aug 30];102(12):4364–74. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022354915308017
25. Chalortham N, Ruangrajitpakorn T, Supnithi T, Leesawat P. OXPIRT: Ontology-based eXpert system for Production of a generic Immediate Release Tablet. In: Formulation Tools for Pharmaceutical Development [Internet]. Elsevier; 2013 [cited 2025 Aug 30]. p. 203–28. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9781907568992500086
26. Lusci A, Pollastri G, Baldi P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. J Chem Inf Model [Internet]. 2013 July 22 [cited 2025 Aug 30];53(7):1563–75. Available from: https://pubs.acs.org/doi/10.1021/ci400187y
27. Hughes TB, Miller GP, Swamidass SJ. Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network. ACS Cent Sci [Internet]. 2015 July 22 [cited 2025 Aug 30];1(4):168–80. Available from: https://pubs.acs.org/doi/10.1021/acscentsci.5b00131
28. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep Learning for Drug-Induced Liver Injury. J Chem Inf Model [Internet]. 2015 Oct 26 [cited 2025 Aug 30];55(10):2085–93. Available from: https://pubs.acs.org/doi/10.1021/acs.jcim.5b00238
29. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B [Internet]. 2019 Jan [cited 2025 Apr 7];9(1):177–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S221138351830282X